Quelques aspects de l'analyse des donnees symboliques

Edwin Diday

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Quelques aspects de l'analyse des données symboliques

Edwin DIDAY

N° 1937
Août 1993

PROGRAMME 5

Quelques aspects de l'Analyse des Données Symboliques[1]
Some aspects of symbolic Data Analysis

Edwin DIDAY
INRIA-Rocquencourt
Domaine de Voluceau
78153 Le Chesnay Cedex

Résumé

Savoir représenter nos connaissances par des expressions à la fois symboliques et numériques, savoir manipuler et utiliser ces expressions dans le but d'aider à décider, de mieux analyser, synthétiser et organiser notre expérience et nos observations, tel est l'objectif que s'assigne l'Analyse des Données Symboliques. On présente d'abord les "objets symboliques" (sortes "d'atomes de connaissances") et ce qui les distingue des objets classiques de l'analyse des données usuelles. Ces objets, qui constituent les individus de l'analyse des données symboliques, permettent de représenter des individus complexes ou des classes d'individus par des conjonctions de propriétés où des descripteurs peuvent prendre des valeurs multiples et pondérées (selon différentes sémantiques) et sont parfois reliés entre eux par des relations d'ordre logique. On introduit des outils pour manipuler ces objets : union, intersection, généralisation, extension etc. On s'intéresse en particulier aux objets symboliques dits probabilistes et l'on énonce un résultat permettant d'étendre les probabilités à ce type d'objet ; on construit ainsi un espace d'objets symboliques probabiliste dual où les individus sont des objets définis en intension, dans cet espace, on pose des problèmes de décomposition en lois de lois. On s'intéresse ensuite à la représentation graphique de ces objets par différentes catégories (hierarchies, pyramides, treillis etc., d'objets symboliques). En utilisant la dualité, on peut construire des suites d'objets symboliques (devenant individus dans l'espace dual suivant) ; ces suites définissent des fractals dans certains cas que nous préciserons. On décrit différents types d'analyse des données symboliques ainsi que les principales étapes d'une telle analyse. On illustre enfin par une application concernant la construction et l'étude de scénarios d'accidents de la route.

Abstract

The main aim of the symbolic approach in Data Analysis is to extend problems, methods and algorithms used on standard data to more complex data called "symbolic objects",in order to distinguish them from objects (described by numerical or categorical variables) treated by standard Data Analysis methods."Symbolic objects" extend classical objects of data analysis in two ways : first, in case of individuals, by giving the possibility of introducing in their definition, structured information ; second, in case of sets or classes, by being intentionally defined.In both cases,in order to represent uncertainty knowledge,it may be useful to use probabilities , possibilities (in case of vagueness and imprecision for instance), belief (in case of probabilities only known on parts and to express ignorance);that why,we introduce several kinds of symbolic objects : boolean, possibilist, probabilist and belief. We focus, in this short paper, on probabilist objects ; a theorem shows how Probability theory may be extended on these objects.Some mixture decomposition problems, on these objects, are settled.We show that in some cases,fractals are well adapted to represent duality between symbolic objects.Sets of symbolic objects are represented by categories of different kinds (hierarchies,pyramids and lattices). Finally,four kinds of data analysis and important steps of a symbolic data analysis are described and illustrated by an example concerning road accidents.

Key-words : Knowledge Analysis, Symbolic Data Analysis, Metadata, Metaknowledge, Probability, Possibility, Evidence theory, Uncertainty Logic,Cognition.

---

[1] Présenté aux journées MODULAD de Lannion (Juin 93)

1. Objectif de l'analyse des données symboliques

Il s'agit d'étendre la problématique, les méthodes et algorithmes de l'analyse des données classique à des données plus riches, car elles expriment une expérience représentant un niveau de connaissance plus élevé, que celle qui est fournie par de simples observations exprimables par un vecteur de valeurs quantitatives ou qualitatives.

De façon générale, il s'agit de se donner la possibilité d'utiliser en entrée des données et des connaissances exprimées par des "objets symboliques", sans craindre de sortir du carcant tabulaire à n lignes (les individus) et p colonnes (les variables) et en évitant de perdre de l'information par des modélisations ou codages arbitraires. Il en est de même en sortie où l'on s'efforcera d'exprimer les résultats par des expressions (sous forme "d'objets symboliques") obtenues automatiquement et possédant un grand pouvoir explicatif par elles-mêmes. Les "objets symboliques", qu'ils soient utilisés en entrée comme en sortie, constituent la clé de voûte de l'analyse des données symboliques. Qu'est-ce qu'un "objet symbolique" ? On peut dire, de façon générale, que c'est une application (définie sur une population observable Ω) munie d'une façon de la calculer pouvant être explicitée à l'aide d'une "description" ; il faut bien distinguer le graphe qui définit l'application (i.e. l'ensemble de ses couples (x,y)) de la façon de la calculer ; considérons par exemple, un ensemble Ω d'élèves d'une classe et deux applications f_1 : $f_1(w) = 1$ si w est grand et 0 sinon $f_2(w) = 1$ si w est roux et 0 sinon ; si dans la classe tous ceux qui sont grands sont roux alors f_1 et f_2 sont identiques alors qu'elles sont construites par des descriptions différentes. Cette description n'est pas nécessairement d'ordre numérique. Ainsi, quand on écrit pour définir une parabole que $f(x) = 2x^2+3$ on donne deux informations : l'une, indique que f est un graphe et l'autre que f(x) est calculé par $2x^2+3$. La description "$2x^2+3$" est purement numérique car elle ne fait intervenir que des opérations numériques portant sur des nombres ou des éléments génériques des nombres ("x" ici). Par contre, les objets symboliques peuvent être décrits par des expressions faisant appel à une axiomatique traduisant une sémantique propre au domaine d'application car l'axiomatique des nombres et des ensembles usuels ne suffit pas ; cela s'avère indispensable quand il s'agit par exemple de décrire et combiner des métadonnées (données sur les données) exprimant à la fois : des contraintes d'ordre, par exemple, logique ("si la couleur est claire, la taille est petite"), de la variation (les fruits produits par cette ferme ont une taille comprise entre 100 et 120 gr.), du doute ("cet avion est civil ou militaire"), du vague ("cet objet est épais et grand"), de l'incertain ("il semble que cet objet soit un avion") etc. Remarquons que ces différentes notions peuvent s'exprimer dans une même description comportant aussi des aspects statistiques ("dans cette espèce de coléoptères il se produit *souvent* qu'un article du tarse soit très petit et passe inaperçu").

2. Ce qui distingue les "objets symboliques" des objets classiques de l'AD.

Les objets symboliques que nous avons étudié (voir Diday 91,92) s'expriment sous forme de "conjonctions" (parfois purement logiques) de propriétés exprimées par des variables classiques (appelés aussi "descripteurs") de l'analyse des données qu'elles soient quantitatives ou qualitatives (nominales ou ordinales). Ils se distinguent des objets classiques de l'AD car il doivent satisfaire aux qualités suivantes :

Les valeurs prises par les descripteurs peuvent être multivaluées afin soit, d'exprimer des classes définies en intension exprimant des variations , soit des individus exprimant du doute.

Plus formellement, un exemple classique d'objet symbolique utilisé en entrée ou en sortie est "l'assertion". Une assertion booléenne se définit comme l'application a : $\Omega \rightarrow$ {vrai, faux} telle que a (w) = $\underset{i}{\wedge}$ [y_i (w) $\in$ V_i] où les y_i sont des fonctions $\Omega \rightarrow O_i$ qui décrivent les éléments de Ω et $V_i \subseteq O_i$. Par convention de notation on écrira a = $\underset{i}{\wedge}$ [$y_i = V_i$]. De cette façon on peut exprimer en intension, une classe d'éléments appelés aussi instances ou individus de Ω, par des propriétés qui les caractérisent.Si a (w) = vrai on dira que w fait partie de "l'extension" de a.

Exemple :
cèpe = [couleur = {jaune, marron}] $\wedge$[taille = [0,15]].
Si w est un cèpe que j'ai trouvé cèpe (w) = vrai si w est soit jaune soit marron et si sa taille est comprise entre 0 et 15 centimètres.

Si maintenant je veux décrire un champignon w pour lequel j'hésite, en ce qui concerne sa couleur, entre blanc et jaune, je peux l'écrire sous la forme :

w^S = [couleur = {blanc, jaune}] $\wedge$ [taille=7.8]

On se trouve dans un tel cas chaque fois, que suite à une observation, une personne exprime un doute entre plusieurs modalités de réponse.

• *Des liens connus entre valeurs prises par les descripteurs doivent pouvoir s'exprimer.*

S'il y a des liens entre les V_i (autrement dit, si les V_i sont fonctions des V_j) cela doit apparaître dans la description de l'objet.

Exemple : la taille des cèpes dépend de la couleur : ils sont clairs quand ils sont petits. On écrit alors

a = [couleur ={jaune, marron}] $\wedge$ [taille = [0,7] si jaune,]7,15] si marron]

Certains descripteurs peuvent ne pas avoir de sens quand d'autres prennent certaines valeurs.

Exemple : quand le descripteur y_1 = "existence d'un chapeau" prend la valeur "non", le descripteur y_2 = "couleur du chapeau" n'a pas de sens. On écrit alors : a = [y_1 = {oui, non}] $\wedge$ [y_2 = jaune, marron, $\emptyset$ si [y_1 = non]].

A l'opposé certains descripteurs peuvent prendre une valeur quelconque dans une description.

Exemple : dans une panne d'un certain type qui a lieu dans un intervalle de température [t=[20,25]] la vitesse du véhicule v : $\Omega \rightarrow$ O n'intervient pas. On écrira alors :

panne = [t = [20,25]] $\wedge$ [v=O] pour exprimer le fait que v peut prendre n'importe quelle valeur dans O, l'ensemble des vitesses possibles.

• *Des liens connus entre parties d'un objet doivent pouvoir s'exprimer*

Dans ce cas, on a étendu la notion d'assertion à celle de "horde" ; une horde est un cas particulier d'objet dit "structuré" en IA ; c'est une application h : $\Omega^p \rightarrow$ {vrai, faux} telle

que si $u = (u_1,...,u_p) \in \Omega^p$ alors $h(u) = \bigwedge_i [y_i(u_i) \in V_i]$; par convention de notation on note $h = \bigwedge_i [y_i(u_i) = V_i]$.

Exemple :

Description d'un scénario d'accident, en faisant intervenir un descripteur y_1 = "type de route", y_2 = "dans", y_3 = "position", les u_i sont des routes et $u = (u_1,...,u_p)$:
scenario $h(u) = [y_1(u_1) = A,B] \wedge [y_1(u_2) = B,C] \wedge [y_2(u_1) = \text{voiture}] \wedge [y_2(u_2) = \text{moto}] \wedge [y_3(u_1,u_2) = \text{croisement multiple}]$.

Autrement dit, ce scénario décrit les accidents qui se produisent entre une voiture dans des routes de type A ou B et une moto dans des routes de type B ou C à un endroit où elles se coupent à un croisement multiple.

• *On doit pouvoir exprimer des propriétés concernant des classes d'individus à l'aide d'expressions relevant de la logique du 1er ordre : les "objets de classes".*

Dans ce cas, l'élément générique qui est décrit est une partie notée C de Ω ; contrairement aux cas précédents où les objets symboliques sont définis sur Ω et décrivent les propriétés d'un individu générique d'une classe, ici on a besoin d'utiliser les quantificateurs $\forall$ et $\exists$ de la logique du premier ordre. On note $P(\Omega)$ l'ensemble des parties de Ω.

Exemple : Un expert en productique (en l'occurence, De Guio de l'ENSAIS à Strasbourg) nous indique que les classes cherchées parmi un ensemble de pièces d'usinage Ω, doivent être d'une part constantes pour une variable y_1 et d'autre part, que les valeurs 1 ou 4 de y_2 et 3 pour y_1 ne peuvent cohabiter ensemble ; on décrira alors ce type de classe sous la forme de l'objet symbolique de classe suivant :
$a : P(\Omega) \rightarrow$ [vrai, faux] tel que
$a(C) = [\forall\, w_1\, w_2 \in C,\ y_1(w_1) = y_1(w_2)] \wedge [\nexists\ w_1,w_2 :$
$y_2(w_2) = [1 \text{ ou } 4]$ et $y_1(w_1) = 3]$.

• *La sémantique et la syntaxe sous-jacente aux données et connaissances d'entrée doit pouvoir s'exprimer : les "objets modaux"*

Les objets symboliques dont il a été question jusqu'ici sont dits "booléens" car il ne prennent que la valeur vraie ou faux (i.e. $a(\omega) \in$ {vrai,faux}) ; dans beaucoup d'applications cela s'avère suffisant, mais souvent l'utilisateur se trouve confronté à des objets où plus de souplesse dans la connaissance qu'il désire exprimer s'avère indispensable. Pour cela, nous avons introduit les objets dits "modaux" car ils "modèrent" les valeurs prises par les variables ; par exemple, un expert pourra dire que la couleur d'un objet d'une classe est souvent rouge et rarement jaune sous la forme de l'objet symbolique a = [couleur=*souvent* rouge, *rarement* jaune] ; a est alors une application de Ω dans l'intervalle [0,1] et $a(\omega)$ exprime un degré de "certitude" quant à l'appartenance de ω à la classe décrite par a . Plus généralement, on peut définir les objets modaux sous la forme : $a = \bigwedge_i x\ [y_i = q_i]$ où q_i peut être une mesure de probabilité, de "possibilité" ou de "crédibilité" (voir Diday* 1992) satisfaisant respectivement aux axiomes de la théorie des probabilités, possibilités (Dubois et Prades 1988) ou des crédibilités (Schafer 1976).

3. Importance du sujet

Avec l'importance grandissante des langages et bases de données orientés objets, les utilisateurs seront de plus en plus amenés à représenter leurs observations et leurs connaissances sous forme d'objets complexes à la fois symboliques et numériques représentant des instances ou définis en intension pour représenter des classes. Le problème de l'analyse et du traitement statistique de telles bases va donc se poser partout avec une acuité grandissante ; au moins dans tous les domaines où il faut représenter, analyser, synthétiser, acquérir et utiliser des connaissances dans un processus automatique.

4. Des outils pour l'analyse des données symboliques

On est amené à définir toute une série d'opérateurs adaptés à la sémantique (notée x) du domaine étudié : $\wedge_x$, $\vee_x$, $\cup_x$, $\cap_x$, C_x (le complémentaire) etc.. Ces opérateurs permettent de généraliser, spécialiser, mesurer des objets symboliques. Un objet a est dit plus général qu'un objet b si l'extension de a contient celle de b. On cherchera alors des opérateurs d'union tels que a $\cup_x$ b soit plus général que a et que b. Mesurer dans un espace d'objets symboliques revient à étendre, par exemple dans le cas probabiliste, les mesures de probabilité à des ensembles d'objets symboliques munis d'opérateurs d'union et d'intersection adéquats (voir Diday 1992) de façon à retrouver l'analogue des axiomes de Kolmogorov sur de tels objets. Cette extension a été aussi réalisée afin de répondre aux cas où les utilisateurs expriment des connaissances convenant à l'axiomatique associée à la théorie des possibilités (Zadeh (1971), Dubois et Prade (1988)) ou celle de "l'évidence" (Choquet 1953, Dempster, Schafer 1990) ; les "possibilités" expriment une sémantique différente de celle des probabilités ; par exemple, pour un dé standard la probabilité d'avoir le 3 est 1/6 alors que sa possibilité est 1 ; si le dé est pipé et si l'on assimile la probabilité du 3 à sa fréquence, on peut avoir une une probabilité nulle bien que sa possibilité puisse être considérée, par un expert, comme faible mais positive ; elles sont bien adaptées par exemple pour la combinaison algébrique de notions "vagues" ("cet objet est *grand*"). La théorie de l'évidence développe une axiomatique propre à traiter entre autres, le cas où l'on ne dispose que de probabilités sur des ensembles de singletons et non sur les singletons eux-mêmes. Nous avons défini des objets symboliques probabilistes, possibilistes et crédibilistes dans Diday (1992) puis nous avons énoncé trois théorèmes montrant que ces objets peuvent être étendus pour exprimer des métaconnaissances (connaissances sur les connaissances) ; plus précisément, si nous disons que les ensembles classiques représentent des connaissances de niveau 0, les probabilités, possibilités et crédibilités des connaissances de niveau 1 (en fournissant des mesures portant sur des combinaisons algébriques d'ensembles de niveau 0), ces trois théorèmes montrent qu'il existe aussi un niveau 2 où l'on peut obtenir des mesures sur des combinaisons algébriques d'ensembles de niveau 1 et satisfaisant à des propriétés analogues à ceux des probabilités, possibilités et crédibilités ; ainsi, au niveau 2 on obtient des sortes de probabilités de probabilités, de possibilités de possibilités et de croyances de croyances.

5. Objets symboliques et outils de traitement

5.1. Objets symboliques et extension

On trouve la définition formelle des objets symboliques, par exemple, dans (Diday 93) ; ici, nous serons plus intuitifs ; disons qu'un objet symbolique considéré dans le cas des "assertions", peut s'écrire en général sous la forme $a = \underset{i}{\wedge_x} [y_i = q_i]$ où a est une application de Ω dans [0,1] qui décrit une partie de Ω sous forme d'une conjonction de

propriétés exprimées par $[y_i = q_i]$; dans le cas particulier d'un objet booleen (voir la description d'un cèpe en 2)) $q_i \in Q_i$ est une fonction caractéristique définie sur un ensemble d'observation noté O_i et y_i est une application de Ω dans Q_i, l'ensemble des applications de O_i dans [0,1] ; ainsi, si y_i désigne la "couleur" et w est un cèpe que j'ai dans la main $y_i(w)$ est la fonction caractéristique r_i associée à la couleur de ce cèpe ; s'il est jaune r_i prendra la valeur 1 pour jaune et 0 ailleurs ; ainsi, $r_i \in Q_i$ est définie sur l'ensemble O_i des couleurs possibles ; w peut lui-même s'exprimer sous forme d'un objet symbolique noté $w^s = \underset{i}{\wedge} [y = r_i]$; par définition, on calcule a(w) par $a(w) = f(\{g(q_i, r_i)\})$ où par exemple, $g(q_i, r_i) = 1$ ssi $q_i(v)$ et $r_i(v)$ sont non nuls simultanément pour tout $v \in O_i$ et $a(w) = f(\{g(q_i,r_i)\}) =$ vrai ssi $\forall i \ \ g(q_i, r_i) = 1$; autrement dit, $f(\{L_i\}) =$ vrai ssi $\forall i$, $L_i =$ vrai; une autre façon de définir f consisterait à dire que a(w) = vrai, quand une proportion suffisament grande de $g(q_i, r_i)$ vaut 1.

L'extension de a est l'ensemble des couples (a,a(w)) ; on peut dire aussi, à un seuil donné α que c'est l'ensemble des w : $a(w) \geq \alpha$. Dans le cas booléan $\alpha=1$

Exemple :

Soit	cepe	= [couleur = {jaune,marron}] $\wedge$ [taille = [0,15]]
Si	$cepe_1$	= [couleur = jaune] $\wedge$ [taille = 16]
et	$cepe_2$	= [couleur = marron] $\wedge$ [taille = 14]

alors, avec le premier choix de f, et si $g(q_i\, r_i) = \sum\{q_i(v)\, r_i(v)/v \in O_i\}$, $cèpe_2$ est dans l'extension de cèpe au seuil $\alpha=1$ mais pas $cèpe_1$.

5.2. Opérateurs d'union, d'intersection et de complémentarité entre objets symboliques

Etant donné $a_1 = \underset{i}{\wedge_x} [y_i = q_i^1]$ et $a_2 = \underset{i}{\wedge_x} [y_i = q_i^2]$ on pose

$a_1 *_x a_2 = \underset{i}{\wedge_x} [y_i = q_i^1 *_x q_i^2]$ où $*_x \in \{\cup_x, \cap_x\}$ et $\bar{a}_1 = \underset{i}{\wedge_x} [y_i = \bar{q}_i]$ où la barre désigne le complémentaire et $\bar{q}_i = 1 - q_i$.

Comment calculer, par exemple, dans le cas booléen l'union, l'intersection et le complémentaire des fonctions caractéristiques $q_i \in Q_i^x$; il suffit d'associer respectivement dans chacun de ces cas la fonction caractéristique de l'union, l'intersection et le complémentaire des ou du support correspondant.

Cas des assertions probabilistes et dualité :

Dans ce cas, les assertions s'expriment sous la forme générale $a = \underset{i}{\wedge_{pr}} [y_i = q_i]$ où $q_i(v) \in [0,1]$ $\forall v \in O_i$ et q_i n'est pas nécessairement une probabilité ; dans le cas particulier où w^s est une assertion probabiliste représentant un individu de Ω, on a : $w^s = \underset{i}{\wedge_{pr}} [y_i = r_i]$ où r_i est une mesure de probabilité.

On calcule a(w) par exemple, de la façon suivante :

$a(w) = f(\{g(q_i, r_i)\}_i)$ où f est la moyenne et

$g(q_i, r_i) = \langle q_i, r_i \rangle = \frac{1}{\text{card } O_i} \Sigma \{q_i(v) r_i(v) / v \in O_i\}$.

Les opérateurs d'union d'intersection et de complémentation s'appliquent de la façon suivante :

$q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1 \, q_i^2$ car quand q_i^1 et q_i^2 sont des lois de probabilités $q_i^1 \cup_{pr} q_i^2(v)$ est la probabilité d'avoir v parmi deux individus de O_i tirés de façon indépendante dans chacune des deux populations représentées par q_i^1 et q_i^2.

$q_i^1 \cap_x q_i^2 = q_i^1 \, q_i^2$ car $q_i^1 \cap_x q_i^2(v) = q_i^1(v) \; q_i^2(v)$ est dans les mêmes conditions la probabilité d'avoir simultanément v dans les deux tirages.

$\bar{q}_i = 1 - q_i$, car $\bar{q}_i(v)$ est la probabilité de ne pas obtenir v.

Si l'on note $\mathbf{a}_{pr}$ l'ensemble des assertions probabilistes on peut étendre leur espace de définition à $\mathbf{a}_{pr}$ en transformant $a_1 : \Omega \rightarrow [\bar{0},1]$ en $a_1^* : \mathbf{a}_{pr} \rightarrow [\bar{0},1]$ tel que $a_1^*(a_2) = f(\{g(q_i^1, q_i^2)\}_i)$ où $a_j = \wedge_{pr}^{i} [y_i = q_i^j]$.

On peut alors démontrer (voir Diday 93) le résultat suivant, qui étend les axiomes de Kolmogorov à l'espace dual :

i) $a^*(\mathbf{a}_{pr}) = 1$

ii) $a^*(A_1 \cup_{pr} A_2) = a^*(A_1) + a^*(A_2) - a^*(A_1 \cap_{pr} A_2)$ où $A_i = \cup_{pr} \{a \in A_i\}$.

En 10 nous donnons des exemples d'objets symboliques probabilistes.

5.3. Généralisation d'objets symboliques

On peut définir une relation d'ordre partielle entre objets symboliques en disant que $a_1 \leq a_2$ ssi l'extension de a_1 est contenue dans celle de a_2 ; on dira alors que a_2 est plus général que a_1 ou que a_1 "hérite" des propriétés de a_2 .
Remarquons que l'union probabiliste est généralisante (comme d'ailleurs l'union booléenne) car $q_1 \cup_{pr} q_2 \geq \text{Max}(q_1, q_2)$ puisque $q_1 \cup_{pr} q_2 = q_1 + q_2 - q_1 \, q_2$. Une autre façon plus précise de généraliser plusieurs assertions $a_1, \ldots, a_n$ par $b_1, \ldots, b_k$, $k < n$ assertions est obtenue en posant le problème suivant : soit $b^* = \cup_{pr}^{j=1,k} b_j^*$, maximiser W tel que $W(b) = \prod_{i=1,n} b^*(a_i)$; on peut aussi choisir $W(b) = \text{Min}_{i=1,n} \, b^*(a_i)$ alors, maximiser W revient à chercher k assertions b_j dont l'extension contienne les $a_1, \ldots a_n$ au seuil le plus grand possible. Dans les deux cas une contrainte à satisfaire peut-être $b_i^*(b_j) = 0$.

5.4. Adéquation d'un objet symbolique avec un ensemble par décomposition de mélange de lois de lois

Etant donné un ensemble d'objets symboliques $A = \{a_1, \ldots, a_n\}$ il s'agit de trouver $b^* = p_1 b_1^* + \ldots + p_k b_k^*$ où p_i est proportionnel à l'extension de $b_i \in \mathbf{a}$ dans A et $\Sigma p_i = 1$ tel que le critère $W(b) = \prod_{i=1,n} b^*(a_i)$ soit maximum sous certaines contraintes sur b et

"5-carré" ; ce quintuple représente le département central de la région que nous voulons décrire ; la méthode* consiste à placer dans chacun des 5 carrés représentant une commune de ce département, la densité de probabilité (ou la distribution des villages (selon les deux paramètres indicateurs de qualité) associés à cette commune ; en plus de la commune centrale on représente aussi dans le carré central, le département par la densité moyenne et (ou) l'union généralisante des densités des 5 communes associées à chaque variable.

Etape p : on note $p^* = **\ldots*$, p fois.

$a^{(p-1)^*}$ est l'ensemble des assertions dont l'extension est un quintuplet de 5^p carrés ; a^{p^*} = $[y^{p^*} = q^{p^*}] \wedge [\text{méthode}^{p^*}]$ où $y^{p^*} : a^{(p-1)^*} \rightarrow Q^{p^*}$ = {ensemble de figures géométriques, composées p fois) ;

$q^* : a^{(p-1)^*} \rightarrow$ {vrai, faux} est tel que $q^{p^*}(w)$ = vrai ssi l'extension de a est un quadruplet de 5^p-carrés de centre (0,0), (n^p,0), (0,n^p), (-n^p,0), (0,-n^p) : la méthode p* consiste à placer dans chaque carré la densité de probabilité (ou la distribution) des villages associées à la commune correspondante ; on ajoute de plus à chaque centre de quintuplet la densité de probabilité ou l'union généralisante des densités de probabilité du département de la région du pays etc. suivant que c'est un centre de département, région ou pays.

Remarquons que l'on pourrait utiliser le carré central pour exprimer les densités moyennes ou généralisantes dans le cas où régions et départements seraient découpés en 4.

La représentation fractale est une description idéale car elle situe communes, départements, régions à égale distance de leur centre (ou "prototype") ; on peut donc l'utiliser pour détecter des anomalies communales, départementales ou régionales en positionnant, carrés "5-carrés" ou 5 fois "5-carrés" à une distance plus ou moins grande (selon leur axe) du centre ; cette distance mesurant l'écart entre la densité de probabilité moyenne ou généralisante d'une commune, d'un département ou d'une région par rapport à son centre.

Chaque case, de la figure 1, contient la description (symbolique, factorielle, probabiliste, catégorielle etc.) de la région ou du département (quand il en est le centre) et de la commune qui lui correspond ; en a) le fractal idéal, en b) le fractal réel avec anomalies, d'une commune (en haut à gauche) et d'un département (en bas). On peut, par exemple, représenter ainsi, en statistiques officielles : une région, cinq départements, 25 communes, 1000 villages.

Remarquons qu'à chaque extrémité du fractal représenté en a), on peut agglutiner quatre nouvelles régions puis aux extrémités ainsi obtenues 4 pays etc. ; signalons aussi que la circularité serait perdue dans le cas d'une représentation en arbre (hiérarchie, ...).

Le fractal ainsi construit est très simple puisque sa forme de base s'écrit a = [y = q] $\wedge$ "méthode" où q est une fonction caractéristique ; on peut obtenir des fractals probabilistes, possibilistes, crédibilistes etc. de plus en augmentant le nombre de propriétés de l'assertion on obtient des fractals basés sur des descriptions de la forme $a = \underset{i}{\wedge}_x [y_i = q_i] \wedge [\text{méthode (i)}]$ pouvant décrire des processus fractals riches comprenant, par exemple, à la fois des descriptions symboliques et des descriptions numériques. Ainsi pour revenir à l'exemple, on aurait pu ajouter à l'évènement qui décrit la position des carrés, un évènement décrivant la densité de probabilité associée à des paramètres mesurant la qualité de l'eau.

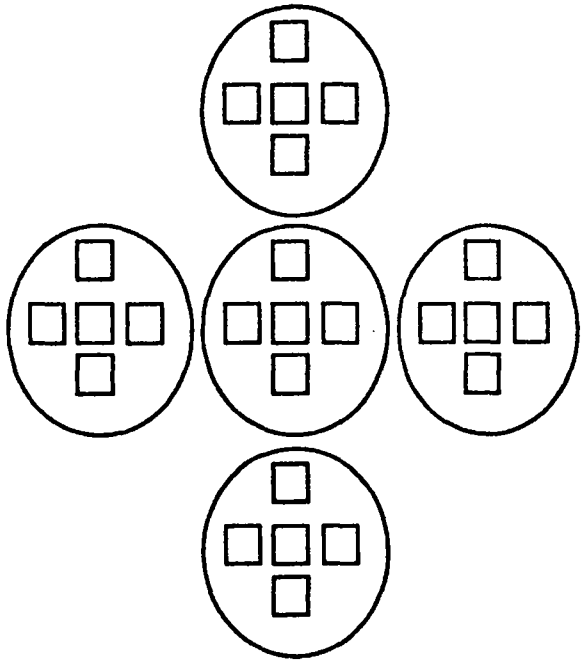

Figure (a)

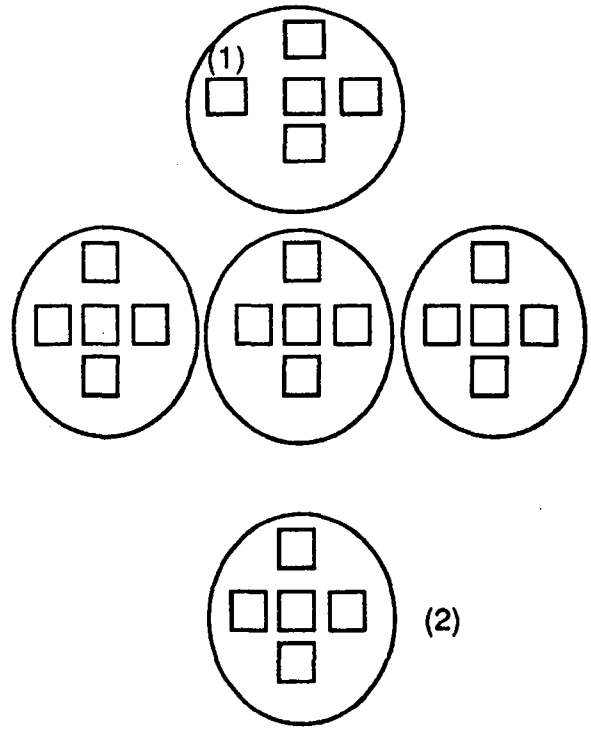

Figure (b)

Figure 1

Un fractal de dimension donnée p est défini par une séquence d'assertions A = (a, a*, ..., a^{p*}) ; on peut considérer qu'il s'agit d'un objet symbolique dont l'extension est définie par l'ensemble des couples (b, A(b)) où b = (w, b, ..., $b^{(p-1)*}$) ∈ (Ω, a, ..., a^{p*}) et A(b) = (a(w), a*(b), ..., a^{p*} ($b^{(p-1)*}$).

Une extension de A à un seuil donné α peut être définie par l'ensemble

B = {b/Min (a(w), a*(b), ..., $a^{p*}(b^{(p-1)*})$) > α). On peut de même, généraliser deux fractals en prenant par exemple leur union symbolique : $A_1 \cup_x A_2 = (a_1 \cup_x a_2, a_1^* \cup_x a_2^*, \ldots, a_1^{p*} \cup_x a_2^{p*})$.

Description symbolique d'une analyse factorielle à l'aide d'une représentation fractale

On peut s'inspirer du fractal de la figure 1 (a) pour représenter une analyse factorielle à l'aide des objets symboliques caractéristiques des extrémités de chaque axe, de la façon suivante : le "département" central représente le plan factoriel 1, 2 ; sa case centrale est l'objet symbolique qui décrit les individus les plus contributifs de ce plan ; ses cases de gauche, de droite, au-dessous et au-dessus décrivant respectivement l'objet symbolique représentant l'extrémité négative et positive de l'axe 1 puis de l'axe 2. Les départements qui se trouvent à gauche à droite, au-dessous et en dessus du département représentent respectivement les plans (1,3), (1,4), (2,3), (2,4) ; on obtient ainsi une "région" d'analyse factorielle ; s'il reste assez d'inertie sur les axes on peut augmenter le fractal d'une dimension et obtenir "un pays" de régions ; la région n° i représentant les différents plans selon la description donnée en figure 2.

Figure 2 (a) Modèle de base pour la représentation de plans factoriels par des objets symboliques ; (b) le choix i = j = 1 pour représenter la première région.

7. Les quatre types d'Analyse des Données

On peut grossièrement définir quatre types d'analyse de données Diday (1987) dont les frontières ne sont pas nécessairement clairement séparées.

(a) "L'analyse des données" classique : on traite des données quantitatives ou qualitatives avec des méthodes numériques utilisant l'algèbre linéaire et les outils de la statistique.

(b) "L'analyse numérique des données symboliques" : on utilise, par exemple, des distances entre objets pour faire une classification ou une analyse factorielle.

(c) "L'analyse symbolique des données classiques". Il s'agit de traiter des tableaux de données classiques (objets caractérisés par des variables quantitatives et (ou) qualitatives) par l'approche symbolique (en utilisant : extensions, ordre symbolique, généralisation, héritage, qualité des objets etc...) soit dés le départ sur les données soit après avoir utilisé une méthode de l'analyse des données classique (afin d'automatiser l'interprétation , par exemple).

Exemple d'analyse symbolique de données classiques :

> Extraire les variables les plus "explicatives" d'un axe factoriel et les 2 classes d'individus les plus contributifs de chaque extrêmité de l'axe. Considérer l'ensembles des objets symboliques associés à ces variables. Trouver dans cet ensemble des objets symboliques complets et d'effritement minimum caractéristiques de chacune des classes. Trouver les objets de meilleure stabilité qui minimisent le recouvrement de la partition associée à ces classes.

(d) "L'analyse symbolique des données" on utilise l'approche symbolique pour traiter des données qui sont aussi symboliques.

Comment se situe l'analyse des données symboliques (ADS) par rapport à d'autres disciplines ?

En statistique, l'introduction d'objets symboliques élargit le champ d'application à des populations qui sont généralement étudiées sous forme de points de $\mathbb{R}^p$ à des populations qui peuvent être formées d'objets complexes exprimant des sémantiques munies d'opérateurs non nécessairement numériques.

En IA, on se situe en amont des systèmes experts puisqu'il s'agit plutôt d'analyser des bases de connaissances ou de les induire à partir des données plutôt que d'étudier des inférences à partir de règles connues. En apprentissage, les problèmes traités sont souvent proches de ceux de l'analyse des données mais s'en distinguent par les objets traités et les méthodes. Ainsi le objets modaux (probabilistes, possibilistes, crédibilistes), par exemple, ont été peu utilisés en entrée comme en sortie des algorithmes d'apprentissage et des méthodes courantes en analyse des données telles que l'analyse factorielle sont négligées.

. Par rapport au "floue" notre objectif se situe à un autre niveau puisque notre but est de faire de l'analyse des données. Peut-on utiliser des fonctions floues pour définir les objets symboliques ? La réponse est affirmitative, par exemple dans le cas particulier des objets possibilistes ; d'autres axiomes que ceux du floues sont, bien sûr, utilisés dans le cas des objets probabilistes et crédibilistes.

8. Les six étapes d'une Analyse des Données Symboliques

On peut grossièrement caractériser l'analyse des données symboliques par six étapes :

1) partir d'un ensemble d'objets individuels plus ou moins complexes ;
2) en extraire des classes par des classifications, analyse factorielle, arbres de décisions, treillis etc;
3) représenter ces classes afin d'obtenir des objets définis en intension sous forme de descriptions (qui peuvent aussi être fournis directement par les experts en court-circuitant ainsi les étape 1) et 2) ;
4) construire à l'aide de ces descriptions des objets symboliques permettant d'identifier des objets individuels;
5) analyser synthétiser, classifier, discriminer, organiser par différentes méthodes d'ADS l'ensemble des objets symboliques de l'étape 4) ;
6) Extraire de ces analyses des métaconnaissances comme par exemple, une pyramide d'héritage et des règles entre objets symboliques que l'on peut en déduire.

9. Exemples d'application

. Scénarios d'accidents : un outil pour les diagnostics de sécurité

Afin d'améliorer la sécurité routière, les spécialistes disposent de grandes banques de données dont les procès verbaux d'accidents rédigés par des gendarmes sur le terrain constituent l'information la plus fine ; à partir des données obtenues en Eure-et-Loire trois experts de l'INRETS, D. Fleury, C.Fline et J.F. Peytavin (1991) ont conçu l'idée de "scénarios d'accidents" afin d'aider à la mise en oeuvre des mesures appropriées pour éviter leur reproduction ; les scénarios définis par les experts sous forme de phrases, décrivent des caractéristiques d'usagers, de déplacements, de lieux de période et parfois de véhicules.

Exemple de scénario d'accident :

"homme de 30-50 ans perdant le controle de son véhicule (usager local, expérimenté, souvent alcoolisé), accident survenant le jour"

Dans un mémoire de DEA MAI réalisé à l'université Paris 9 Dauphine principalement sous la direction de D. Fleury (INRETS) et M. Summa (Lise-Ceremade), A. Regnier a appliqué plusieurs étapes de l'analyse des données symboliques afin d'améliorer, compléter et organiser la base de scénarios d'accidents fournie par les experts.

Dans ce travail chaque scénario est conçu comme "intension" d'une classe d'accidents de la base de données des gendarmes ; les principales étapes de l'étude ont été les suivantes :

a) exprimer les scénarios sous forme d'objets symboliques ; ce sont les objets probabilistes qui se sont avérés les plus adéquats.

Les programmes épars existent déjà (pyramides symboliques, générateurs automatiques de règles, arbres de segmentation sur objets symboliques etc...). Un défi à relever va être maintenant de concrétiser toutes ces avancées par un grand logiciel soit sous forme de bibliothèque de programmes (une sorte de MODULAD symbolique) soit sous forme plus ambitieuse d'un système interactif (sorte de SICLA symbolique).

Références

P. Brito, E. Diday (1990), "*Pyramidal Representation of Symbolic Objects*", in knowledge data and computer assisted decision. M. Schader, W. Gaul edi NATO ASI Series.

G. Choquet, (1953), "*Théorie des capacités*", Ann. Inst. Fourrier 5. 131-295.

A.P. Dempster, (1966), "*Upper and Lower Probabilites generated by random closed interval* ", Annals of Mathematical Statistics 39, 957-966.

E. Diday (1987), " *Introduction à l'approche symbolique en analyse des données* ", Journées symboliques "Symbolique-Numérique" pour l'apprentissage de connaissances à partir de données. E. Diday, Y. Kodratoff, éditeurs LISE-CEREMADE.

E. Diday (1991), "*Towards a statistics of intension for knowledge analysis*", WocFac'91 Paris, Angkor edit.

E. Diday (1991), "*From Data Analysis to Uncertainty Knowledge Analysis*" Proc. ECSQAU, Symbolic and Quantitative Approaches to Uncertainty. Marseille, Springer-Verlag.

E. Diday (1991), "*Knowledge Representation and Symbolic Data Analysis* ", in knowledge data and computer assisted decision. M. Scheder, W. Gaul edit NATO ASI Series.

E. Diday (1991), "*Des objets de l'analyse des données à ceux de l'analyse des connaissances* ", in Induction Symbolique et Numérique Y. Kodratoff et E. Diday edi CEPADUES-Edition.

E. Diday (1992), "*Belief Objects*" Proc. IPMU On Information Processing and Management of Uncertainty in knowledge based systems. Univ. des Iles Baléares édit.

E. Diday (1992), "*From Data to Knowledge : Probabilist objects for a symbolic data analysis* ", In Computational Statistics Vol 1, p. 193, Y. Dodge and J. Whittaker Editors. Physica Verlag.

E. Diday (1992), "*Probabilist, Possibilist and Belief Objects for knowledge analysis*" Journées "Symbolique-Numérique" Université Paris 9 Dauphine édité par le Lise-Ceremade. Un texte plus complet à paraître dans "Annals of Operations research est disponible à l'INRIA.

E. Diday (1992), "*Symbolic data analysis*", Tutorial à IFCS'93.

Dubois D., Prade H., (1988), "*Possibility theory* ", Plenum New York.

K.C. Gowda, E. Diday (1991), "*Symbolic clustering using a new dissimilarity measure*", Pattern Recognition, Vol. 24, N° 6.

K.C. Gowda, E. Diday (1991), "*Unsupervised learning through symbolic clustering* ", Pattern Recognition letters, Vol. 12, N° 5.

C. Gowda, E. Diday (1992), "*Symbolic Clustering Using a New Similarity Measure* ", IEEE Tr. on Systems, Man, and Cybernetics. Vol. 22, N° 2.

A. Regnier (1992), "*Analyse Symbolique de scénarios d'accidents*" , Rapport de DEA MASE, Université Paris IX-Dauphine.

G. Schafer, (1990), "*Perspectives on the Theory and Practice of Belief functions* "International Journal of Approximate Reasoning. Vol 4, Numbers 5/6.

L.A. Zadeh (1971) - "*Quantitative fuzzy semantics*". Informations Sciences, 159-176).

Annexe

La théorie des catégories a été introduite en 1945 par Eilenberg et Maclane pour rendre compte de propriétés très générales des structures mathématiques. Ici, on s'inspire de D.E. Rydeheard et R.M. Burstall et surtout B. Mitchell : (Theory of Categories, Academic Press, New York 1965).

Définition

Une catégorie C est constituée par la donnée d'un quadruplet : C = (O,A,s,t) où O et A sont des ensembles et s, t sont deux applications de A dans O. On appelle O, l'ensemble des nœuds, A l'ensemble des flèches ; si s (f) = a, est l'origine de la flèche f∈ A,. b=t (f) est sa terminaison, on note f : a → b. C est une catégorie ssi les conditions suivantes sont satisfaites :

1) ∀ a ∈ O la flèche i_a : a → a existe,

2) ∀ f,g,h ∈ A et a, b, c ∈ O tels que f : a → b, g : b → c et h : c → d on a
(hg) f = h (gf) et f i_a = f = i_b f

Exemples de catégorie

1) Un ordre partiel a ≤ b ≤ c, a ≤ d : la catégorie associée est définie par

O = {a,b,d}, A = l'ensemble des flèches f : a ≤ b <=> f : a → b

On peut représenter cette catégorie par le schéma donné en figure 4.

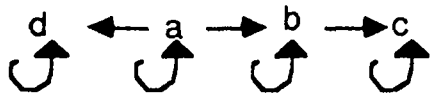

Figure 4 : catégorie d'ordre partiel

2) Un pyramide : O = ensemble des classes et des individus de la pyramide

A = ensemble des flèches f : a ⊆ b <=> f : a → b

3) Le treillis des objets complets : O = les objets complets, A = {f/a ≤ b <=> f : a → b }

Définition d'un foncteur

C'est une paire de fonctions notées F toutes deux entre deux catégories A et B, l'une reliant les objets de A à ceux de B et l'autre reliant les flèches de A à celles de B.

. F : Obj (A) → Obj (B) , F : a → F (a)

. F : Flèches (A) → Fleches (B), F : (f : a → b) => F (f) : F (a) → F(b) satisfaisant :

F (id_a) = $id_{F(a)}$ et F(gf) = F(g) F(f) ∀ gf défini.

Exemple :

1) Le passage d'une pyramide à sa représentation graphique est obtenue à l'aide d'un foncteur.
2) Le passage de l'ensemble des objets symboliques complets à leur représentation graphique sous forme de treillis est obtenu à l'aide d'un foncteur.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine - Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes - IRISA, Campus universitaire de Beaulieu 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 Grenoble Cedex 1 (France)
Unité de recherche INRIA Sophia Antipolis - 2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

ISSN 0249 - 6399